Delta-KNN: Improving Demonstration Selection in In-Context Learning for Alzheimer's Disease Detection

Anonymous ACL submission

Abstract

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that leads to dementia, and early intervention can greatly benefit from analyzing linguistic abnormalities. In this work, we explore the potential of Large Language Models (LLMs) as health assistants for AD diagnosis from patient-generated text using in-context learning (ICL), where tasks are defined through a few input-output examples. Empirical results reveal that conventional ICL methods, such as similarity-based selection, perform poorly for AD diagnosis, likely due to the inherent complexity of this task. To address this, we introduce **Delta-KNN**, a novel demonstration selection strategy that enhances ICL performance. Our method leverages a delta score to assess the relative gains of each training example, coupled with a KNN-based retriever that dynamically selects optimal "representatives" for a given input. Experiments on two AD detection datasets across three open-source LLMs demonstrate that Delta-KNN consistently outperforms existing ICL baselines. Notably, when using the Llama-3.1 model, our approach achieves new state-of-the-art results, surpassing even supervised classifiers.[1]

1 Introduction

Large Language Models (LLMs), powered by advanced deep learning and vast cross-disciplinary training data, have transformed Natural Language Processing (NLP) (Zhao et al., 2023; Fan et al., 2024). They show promise in specialized fields like clinical medicine and healthcare (Bubeck et al., 2023; Cui et al., 2024; Belyaeva et al., 2023; Jin et al., 2024). However, their ability to outperform traditional AI in tasks requiring deep understanding and nuanced analysis remains uncertain (Wang et al., 2023b).

In this paper, we investigate LLM's capabilities in a crucial healthcare challenge: **Alzheimer's Disease** detection. AD is a severe neurodegenerative disorder affecting 55 million people worldwide, ranking among the most costly diseases[2]. Our approach focuses on identifying AD patients based on their descriptions of a given image, such as the well-known *Cookie Theft* picture (Figure 1). Traditional machine learning methods typically rely on feature extraction (e.g., linguistic analysis) (Fraser et al., 2016, 2019; Barral et al., 2020) or embedding-based models (e.g., BERT) (Balagopalan et al., 2021) to convert speech into vectors for classification. However, NLP has shifted from task-specific models to task-agnostic foundation models (Radford et al., 2019; Brown, 2020), enabling LLMs to not only simplify the diagnostic process but also offer interpretable reasoning, providing clinicians with valuable insights into their decision-making (Perlis, 2023; Nori et al., 2023a,b).

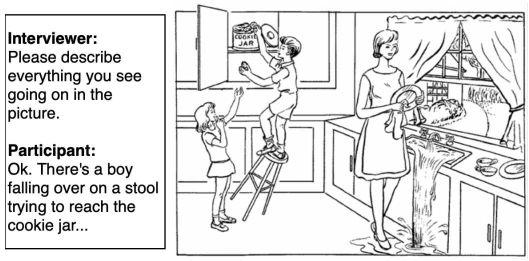

Figure 1: The Cookie Theft picture description task.

A major challenge in leveraging LLMs for AD detection is how to effectively teach them to learn from very **limited data** (i.e., few hundreds examples). In-context learning (ICL)—where a model performs a new task by conditioning on a few input-label pairs during inference—has emerged as a powerful and widely adopted strategy for handling complex tasks, which is applicable in data-poor scenarios. One common approach involves *similarity-based* selection, where examples resembling the target input or output are chosen. This method has

[1] Our code will be released.

[2] https://www.who.int/news-room/fact-sheets/detail/dementia.

shown strong performance in tasks like question answering, commonsense reasoning, and text-to-SQL generation (Liu et al., 2022; Su et al., 2023; Li et al., 2025), but one concern is that the adopted similarity metrics may only capture a shallow understanding of the text. In order to enhance the model's *understanding* of the target sample, Peng et al. (2024) proposed a method that minimizes the conditional entropy between the demonstration and target, demonstrating improvements on both classification and generation tasks.

Other concerns include that ICL is highly sensitive to the selection of demonstration examples (Lu et al., 2022; Iter et al., 2023) and often struggles with tasks requiring complex reasoning (Peng et al., 2023). In light of these limitations, not surprisingly, our preliminary experiments reveal that existing ICL methods perform poorly on AD detection from text, which arguably requires the model to capture very subtle and complex linguistic and conceptual differences.

To address these challenges, we introduce a novel demonstration selection method, denoted as Delta-KNN, that practically quantifies the *expected gain* of each example. This gain, referred to as the **delta score**, measures the improvement in model performance before and after including a demonstration. Using a small held-out set, we construct a **delta matrix** that stores performance gains for all examples. At inference time, we first identify target "representatives" by finding the **nearest neighbors** based on text similarity between the target sample and the held-out examples. Then, we select demonstrations that maximize the *expected gain* for these representatives. Extensive experiments on two AD detection datasets confirm the effectiveness of our approach, consistently outperforming existing demonstration selection methods. Additionally, we evaluate its robustness across different LLMs and explore its synergy with prompt engineering, achieving state-of-the-art (SOTA) performance comparable to supervised baselines.

In summary, (1) we introduce a **novel ICL method** designed to capture complex linguistic and conceptual nuances, making it particularly powerful in data-scarce scenarios; (2) Our approach achieves **state-of-the-art performance**, surpassing existing ICL baselines in detecting dementia, which is one of the most costly diseases worldwide; (3) Through extensive experiments, we show that the benefits of our method are conveniently **model- and prompt-agnostic**.

2 Related Work

Language Analysis for AD detection. Clinical studies have established a strong connection between speech and language abnormalities and AD pathology (Sajjadi et al., 2012; Rodríguez-Aranda et al., 2016). Research in this area mostly relies on data from the Cookie Theft picture description task, particularly from the DementiaBank (Becker et al., 1994) and ADReSS (Luz et al., 2021) datasets, and utilize semantic, syntactic, and lexical features (Ahmed et al., 2013; Fraser et al., 2016, 2019; Jang et al., 2021), with some studies also incorporating information unit analysis, such as counting object mentions in the picture (Masrani et al., 2017; Favaro et al., 2024). While these methods achieve strong performance, they often rely on manual data annotation and feature engineering. To reduce the need for labor-intensive processes, recent studies have explored deep learning approaches, including transfer learning (Zhu et al., 2021; Balagopalan et al., 2021; Agbavor and Liang, 2022), neural networks (Kong et al., 2019; Fritsch et al., 2019; Bouazizi et al., 2023), and LLMs (Achiam et al., 2023). However, given the complexity of AD detection from text, naively prompting LLMs does not yield promising results (Wang et al., 2023b). Instead, more sophisticated in-context learning strategies are required to fully explore LLM's inner specialist capabilities.

Demonstration Selection in ICL. Few-shot in-context learning (ICL) with LLMs has demonstrated performance comparable to supervised fine-tuning across various tasks like reasoning (Wei et al., 2022; Dong et al., 2022). However, its effectiveness remains highly dependent on demonstration selection, leading to instability (Lu et al., 2022; Peng et al., 2023). While Lu et al. (2022) explored the impact of example order, they did not propose a method for selecting better examples. Liu et al. (2022) found that semantically similar examples improve ICL, later extended by incorporating more diverse demonstrations (Su et al., 2023). Other studies have focused on enhancing model understanding through ranking mechanisms (Wu et al., 2023), perplexity-based prompt evaluation (Gonen et al., 2023), and conditional entropy to assess model comprehension (Peng et al., 2024). While these methods perform well on standard benchmarks, they remain untested in tasks like AD detection, where capturing subtle linguistic differences and reasoning-based cues is critical.

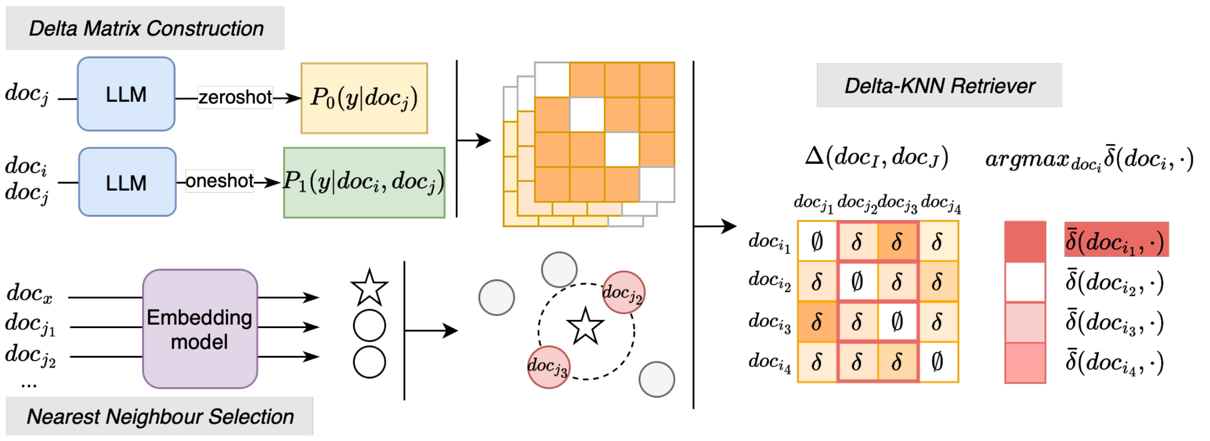

Figure 2: Delta-KNN retriever pipeline consists of two parts: (1) construct a delta matrix Δ by calculating the performance gain from each demonstration example; (2) search for nearest neighbors (e.g., doc_{j_2}, doc_{j_3}) for target doc_x in a vector space. The best demonstration example (doc_{i_1}) maximizes the average delta score over nearest neighbors ($\delta(doc_{i_1}, doc_{j_2}), \delta(doc_{i_1}, doc_{j_3})$).

3 Method

Our demonstration selection method consists of two modules: the first module constructs a *performance gain* matrix using LLMs, referred to as the **Delta Matrix** (Δ). Each cell in Δ contains a **delta score** (δ) to represent the improvement or degradation contributed by a specific demonstration doc_i to a target doc_x. The second module employs an embedding-based retriever, which prioritizes demonstration examples based on their vector similarity to the target example. By integrating the outputs of both modules, we compute the average delta score ($\bar{\delta}$) for the k nearest neighbors of each example in the training set. The optimal demonstration example is identified as the one with the highest aggregated delta score ($argmax(\bar{\delta})$). This process is illustrated in Figure 2, and we describe each module in detail in the following sections.

Delta Matrix Construction with LLM. We construct the Delta Matrix by prompting the LLM in both zero-shot and one-shot scenarios. In the zero-shot scenario, the input to the LLM consists of the text from a document doc_j. To elicit a probability alongside the predicted label, we include the cue phrase "Give a prediction with a probability" in the prompt, which has shown good calibration capabilities (Tian et al., 2023). P_0 is the probability of the correct label for zero-shot prediction: $P_0(\hat{y}|doc_j; \theta)$ where θ refers to LLM parameters.

In the one-shot scenario, the predicted label and probability are obtained by appending the whole example i (text and label) prior to the text of doc_j. Similarly, we obtain the probability of correct prediction in one-shot: $P_1(\hat{y}|doc_i, doc_j; \theta)$.

The **delta score** for a demonstration example doc_i relative to the target doc_j is defined as the difference between P_1 and P_0:

$$\delta(doc_i, doc_j) = P_1(\hat{y}|doc_i, doc_j; \theta) - P_0(\hat{y}|doc_j; \theta) \tag{1}$$

In a training set D with d number of documents, the **Delta Matrix** Δ is a $d \times d$ matrix where each cell $[i, j]$ contains a delta score $\delta(doc_i, doc_j)$, representing the relative gain when doc_i is used as a demonstration for doc_j:

$$\forall doc_i \in D, doc_j \in D, i \neq j, \Delta = \delta(doc_i, doc_j) \tag{2}$$

Similarity-based k Nearest Neighbors (KNN). We hypothesize that the average delta score $\bar{\delta}$ derived from guiding the documents **most similar** to the target document is more informative and effective. Thus, we include a second module to select the most similar documents. Specifically, we use a embedding model to convert documents to vector representations. For each target doc_x, we identify its k nearest neighbors doc_{j_1}, doc_{j_2}, ..., doc_{j_k} from the training set based on the distances in the embedding space. Using predefined similarity metrics, such as cosine similarity, the neighbors are ranked such that $doc_{j_1} < doc_{j_2}$ if $\cos(doc_{j_1}, doc_x) > cos(doc_{j_2}, doc_x)$. Note that the number of neighbors can vary: $k \in [1, d]$. In practice, we conduct simulations across a range of k values within the training set using five-fold cross validation and apply the optimal k value to the test documents.

A key element in this step is the **embedding**

model, which projects sentences into a latent semantic space. We choose OpenAI embeddings (Neelakantan et al., 2022) as text encoder because it has shown superior results on a series of information retrieval tasks (Xian et al., 2024; Lin et al., 2023), surpassing methods such as BM25 (Robertson et al., 2009), leaned sparse representations uniCOIL (Ma et al., 2022), and other semantic embedding APIs (Kamalloo et al., 2023). Precisely, we employ the latest `text-embedding-3-large` model[3]. Additionally, we compare this with representations obtained directly from LLMs (§5.5).

Delta-KNN Retriever. By combining the two modules described above, we compute the average delta score (Equation 1) of each demonstration doc_i over the k *most similar* documents $(doc_{j_1}, ..., doc_{j_k})$ to doc_x:

$$\bar{\delta}(doc_i, \cdot) = \frac{1}{k} \sum_{k'=1}^{k} \delta(doc_i, doc_{j_{k'}}) \tag{3}$$

This delta score represents the *expected* gain when using doc_i as a demonstration to the target doc_x. Mathematically, we aim to find the best doc_i by solving the following optimization problem:

$$doc_{i^*} = \arg\max_{doc_i \in D} \bar{\delta}(doc_i, \cdot) \tag{4}$$

where doc_{i^*} is the example that maximizes the average delta score. In n-shot ICL, we rank the examples in descending order and concatenate them to form the context $\{doc_{i_1}, ...doc_{i_n}\}$ prior to the target doc_x.

The Delta Matrix offers an intuitive map to guide the demonstration selection. Different from existing similarity-based methods (Nori et al., 2023b) or text-understanding-based retrieval approaches (Peng et al., 2024), our method is grounded in empirical evidence of performance gains observed from semantically similar documents.

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

Picture Description Task, such as the one shown in Figure 1, is a widely used task to capture deficits or abnormalities in language (Yorkston and Beukelman, 1980; Favaro et al., 2024). In this work, we use two datasets that contain *Cookie Theft* picture description for AD detection: ADReSS and Canary. **ADReSS** (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge dataset (Luz et al., 2021) is a curated subset of DementiaBank's Pitt Corpus (Becker et al., 1994) that is matched for age and gender. It consists of 156 speech recordings and transcripts from AD (N = 78) and non-AD (N = 78) participants, and is divided into a training set and a test set. **Canary** is collected by Jang et al. (2021), comprising 63 patients recruited from a specialty memory clinic and 67 healthy controls from the community. Patients are either diagnosed with AD or exhibiting initial symptoms of Mild Cognitive Impairments potentially progressing to AD. Canary includes longer documents with greater variations in length, gender, and age compared to ADReSS (see details in Appendix A), making it a more challenging dataset while more accurately reflecting clinically collected data.

For **evaluation metrics**, we use (1) Accuracy (ACC), (2) Area Under the Curve (AUC) which captures the ability to distinguish between Patient and Control under different thresholds, (3) Sensitivity (SEN): the True Positive rate for Patient detection, and (4) Specificity (SPE): the True Negative rate for Control detection.

4.2 Baselines

We compare our approach with popular demonstration selection methods. Since constructing the Delta Matrix relies on information from a training set, we also benchmark with supervised methods.

Demonstration Selection Methods. Including:

(1) Zero-Shot: A special case of ICL where no demonstration example is given.

(2) Random Sampling: Randomly select examples for each target i.

(3) Similarity-based Top-k Selection: Proposed in Liu et al. (2022) and has been widely used for health-related ICL (Nori et al., 2023b,a), where examples are embedded in a vector space and the nearest neighbors (calculated using cosine similarity) are selected as demonstration.

(4) Text-understanding-based CE Selection: A recent approach that quantifies *understanding* by measuring the Conditional Entropy (CE) of the target input given a demonstration and it selects examples that minimize the CE (Peng et al., 2024).

Supervised Baselines. Including:

(1) Statistical Machine Learning Classifiers: Traditional methods that use feature extraction (e.g., lexico-syntactic and semantic features) and supervised algorithms like Support Vector Space

[3]https://openai.com/index/new-embedding-models-and-api-updates/

Method	ADReSS-train				ADReSS-test				Canary			
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE
Zero-shot	$62.2_{0.0}$	$60.1_{0.0}$	$\mathbf{98.1}_{0.0}$	$22.2_{0.0}$	$57.6_{1.0}$	$57.6_{1.0}$	$\mathbf{100.0}_{0.0}$	$15.3_{2.0}$	$73.3_{0.4}$	$72.1_{1.0}$	$79.4_{0.0}$	$67.7_{0.7}$
Random	$68.4_{2.2}$	$71.9_{3.1}$	$84.0_{2.3}$	$48.8_{6.3}$	$75.7_{4.3}$	$81.5_{2.6}$	$93.1_{2.0}$	$58.3_{9.0}$	$73.1_{2.7}$	$75.3_{3.7}$	$72.0_{3.3}$	$74.1_{2.5}$
Top-k Select.	$69.0_{1.6}$	$71.9_{2.5}$	$88.3_{2.3}$	$45.7_{1.7}$	$70.1_{2.0}$	$80.0_{0.8}$	$91.7_{3.4}$	$48.6_{2.0}$	$71.0_{2.5}$	$75.0_{2.2}$	$76.7_{0.7}$	$65.7_{4.2}$
CE* Select.	$67.4_{2.3}$	$74.5_{1.3}$	$85.2_{1.5}$	$45.7_{3.1}$	$70.1_{1.0}$	$76.4_{2.6}$	$93.1_{2.0}$	$47.2_{2.0}$	$73.3_{1.9}$	$78.4_{0.9}$	$\mathbf{79.9}_{2.0}$	$67.2_{4.4}$
Delta-KNN (ours)	$\mathbf{79.2}_{1.2}$	$\mathbf{78.9}_{1.3}$	$69.1_{0.9}$	$\mathbf{85.2}_{1.5}$	$\mathbf{80.5}_{3.9}$	$\mathbf{85.8}_{0.9}$	$70.8_{5.9}$	$\mathbf{86.1}_{2.0}$	$\mathbf{78.5}_{1.5}$	$\mathbf{79.8}_{0.9}$	$70.6_{0.8}$	$\mathbf{85.8}_{2.2}$

Table 1: AD detection results (accuracy, AUC, sensitivity, specificity) on ADReSS train set, ADReSS test set, and Canary using different demonstration selection methods. We compare with zero-shot, random sampling, Top-k (Liu et al., 2022), and CE*-based (conditional entropy) selection (Peng et al., 2024). All results are averaged over three runs with standard deviation in subscription. Best score per column is in **bold**.

(SVM) (Luz et al., 2021), Random Forest (RF) (Luz et al., 2021), Logistic Regression (LR) (Jang et al., 2021), and simple structure Neural Network (NN) (Balagopalan et al., 2021). We replicate a few studies and report results in §5.7 and Appendix B.

(2) Transfer Learning-based Language Models: Pretrained Language Models (PLMs) like BERT (Devlin, 2018) encode rich linguistic information and are often fine-tuned for classification tasks without the need for manual feature extraction. We fine-tune a BERT model by following Balagopalan et al. (2021) (details in Appendix C) and include results from a SVM classifier which uses GPT-3 embeddings for contextualized input (Agbavor and Liang, 2022).

(3) Supervised Fine-tuning: It is a common approach to adapt LLMs for downstream tasks by training on task-specific data, updating some or all parameters. We employ LoRA (Hu et al., 2022), a parameter-efficient fine-tuning strategy.

4.3 Implementation Details

Our experiments are conducted on Llama-3.1-8B-Instruct (Dubey et al., 2024). To assess the robustness of our method, we also test it on Qwen2.5-7B-Instruct (Yang et al., 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), see §5.6.

For zero-shot and few-shot ICL, we use a low temperature (0.01) and set top_k sampling to 50. We use 4-shot learning with two positive and two negative examples. The impact of in-context examples is discussed in §5.3. To address the potential *non-determinism* of LLMs (Ouyang et al., 2023; Song et al., 2024), each experiment is tested three times. We present the average scores along with the standard deviation. For LLM fine-tuning, we employ LoRA technique (Hu et al., 2022) and train the model for one epoch, with details in Appendix D.

Given the complexity of this task (Bouazizi et al., 2023; Favaro et al., 2024), we carefully design prompts with comprehensive instructions to enhance the model's understanding and diagnostic capabilities. Our prompt includes: `Role`—"*You are a medical expert in Alzheimer's Disease*" to establish domain expertise, `Context`—a concise introduction to the Cookie Theft picture description task, and `Linguistic`—key linguistic features the model should focus on. In addition, we incorporate a Guided Chain of Thought (`G.-CoT`) reasoning step (Kojima et al., 2022), prompting the model to analyze specific linguistic aspects such as *vocabulary richness* and *syntactic complexity*, supported by clinical observations (Ash and Grossman, 2015; Forbes-McKay and Venneri, 2005; Bouazizi et al., 2023). A complete prompt template (`Role+Context+Linguistic`; `Demonstrations`; `G.-CoT`) is provided in Appendix E. Preliminary zero-shot experiments validate the effectiveness of this prompt design. To further analyze its impact, we introduce variations by ablating different components and evaluating them in §5.2.

5 Experiments

We conduct experiments to show the effectiveness of Delta-KNN compared with other ICL methods (§5.1) and examine the influence of prompt engineering (§5.2) and hyperparameters (§5.3, §5.4). Further investigation involves using different embedding methods (§5.5) and LLMs (§5.6), with a comparison against supervised classifiers (§5.7).

5.1 Delta-KNN *vs.* Other Demonstration Selection Methods

Table 1 presents AD detection results on the ADReSS and Canary datasets using Random sampling, Top-k, and CE-based selection. Additionally, we report zero-shot results. We prompt LLM with the (`Role+Context+Linguistic`; `Demonstrations`; `G.-CoT`) template, with all ICL methods containing four demonstrations with

	Role	Con.	Ling.	CoT	G.-CoT	ADReSS-train				ADReSS-test				Canary			
						Delta-KNN (Rdm, Top$k$, CE)				Delta-KNN (Rdm, Top$k$, CE)				Delta-KNN (Rdm, Top$k$, CE)			
(1)	✗	✗	✗	✗	✗	73.0	↓ 13.9	↓ 17.6	↓ 12.7	69.8	↑ 0.3	↓ 2.8	↑ 0.8	63.1	↓ 2.3	↓ 2.3	↓ 3.9
(2)	✓	✓	✗	✗	✗	72.7	↓ 2.1	↓ 2.4	↓ 3.3	69.1	∼ 0	∼ 0	↓ 2.9	70.0	↓ 3.8	↓ 3.8	↓ 4.1
(3)	✓	✗	✓	✗	✗	73.1	↓ 7.8	↓ 13.1	↓ 6.9	74.4	↓ 5.3	↓ 2.6	↓ 1.2	68.1	↓ 9.1	↓ 2.2	↓ 2.7
(4)	✓	✓	✗	✓	✗	73.6	↓ 5.9	↓ 4.9	↓ 6.3	74.6	↓ 2.1	↓ 2.8	↓ 2.8	71.5	↓ 9.7	↓ 4.3	↓ 10.2
(5)	✓	✗	✓	✓	✗	74.5	↓ 10.2	↓ 14.5	↓ 16.3	74.6	↓ 11.1	↓ 13.2	↓ 16.7	65.1	↓ 3.6	↓ 6.9	↓ 7.7
(6)	✓	✓	✓	✓	✗	80.0	↓ 9.9	↓ 11.1	↓ 8.6	83.6	↓ 13.8	↓ 10.4	↓ 12.5	70.8	↓ 7.5	↓ 7.7	↓ 9.6
(7)	✓	✓	✓	✗	✓	79.2	↓ 10.8	↓ 10.2	↓ 11.8	80.5	↓ 2.8	↓ 8.4	↓ 8.4	78.5	↓ 5.4	↓ 7.5	↓ 5.2

Table 2: Delta-KNN performance (accuracy) using different prompt engineering strategies (Role, Context, Linguistic cues, chain-of-thought reasoning (CoT), and Guided CoT) on ADReSS and Canary datasets, in comparison with Random sampling (Rdm), Top-k selection (Liu et al., 2022), and Conditional Entropy (CE) (Peng et al., 2024) baselines. ↓, ↑, and ∼ symbols refer to lower, higher, and same accuracies compared to Delta-KNN, respectively.

balanced labels (§5.3). The k value in Delta-KNN is set to 13, which is based on empirical results on training sets (§5.4).

A first interesting observation from Table 1 is that zero-shot prompting almost always predicts participants as Patients, achieving as high as 100% sensitivity while failing to identify Controls, a **bias** particularly evident in the ADReSS dataset. With in-context learning, models exhibit more balanced predictions—most ICL methods significantly improve specificity, with higher accuracy and AUC scores. This shows that learning from examples helps correct the model's initial bias. Random sampling performs well overall, suggesting that exposure to a diverse input distribution benefits ICL (Nori et al., 2023a). Surprisingly, the recent CE-based selection delivers mixed results. While it improves performance on Canary, it falls short on ADReSS compared to Random sampling and Top-k selection. In contrast, our proposed method consistently outperforms all selection methods on both datasets, achieving a $5-10\%$ and 5% accuracy improvement on ADReSS and Canary, respectively. Notably, Delta-KNN excels at identifying speech from healthy controls (SPE: $70-85\%$) while maintaining strong performance in detecting patients (SEN: 70%). Overall, our method attains an optimal AUC score ($79-85\%$), highlighting the strong discriminative power of the selected examples.

5.2 Impact of Prompting Engineering

It is known that a model's performance can be significantly affected by its prompt, often in surprising ways (Feng et al., 2024; Sivarajkumar et al., 2024; Salinas and Morstatter, 2024; Sclar et al., 2024). To examine the impact of prompt engineering and assess the robustness of our approach, we conduct ablation studies on prompt engineering. Precisely, we systematically vary the prompt design by gradually removing task-related information, ranging from a minimal prompt (lacking background details and CoT reasoning cues) to a comprehensive prompt containing all key components, i.e., (`Role+Context+Linguistic`; `G.-CoT`).

In Table 2, we test seven variations using Delta-KNN, comparing it with Random sampling, Top k, and CE-based selection. The results clearly show that task-related information is crucial: without prompt engineering (Prompt 1), Delta-KNN achieves 69% accuracy on ADReSS-test and 63% on Canary, which are 11 and 15 points lower than the best-performing design (Prompt 7). Adding background details such as `Role`, `Context`, and `Linguistic` (Prompts 2 and 3) improves accuracy by 5%, confirming the importance of domain-specific context. When including a simple Chain-of-Thought (`CoT`) cue phrase "*First explain step-by-step and then give a prediction.*", prompts 4 and 5 give further gains. Although marginal, it significantly enhances interpretability by making the model's reasoning more transparent. Finally, combining all background information with `CoT` (Prompt 6) boosts performance, with the highest accuracy achieved using our `Guided CoT` (Prompt 7). Remarkably, across all prompt settings, Delta-KNN consistently outperforms other demonstration selection methods, demonstrating its robustness under different prompting strategies.

5.3 Impact of In-Context Examples

To assess the impact of in-context learning, we gradually increase the number of examples (N) from 0 to 12. As shown in Figure 3, performance generally improves with more examples.

Interestingly, when using only one example, most selection methods experience a sharp performance drop compared to zero-shot, likely due to biased label distribution in demonstrations (Min et al.,

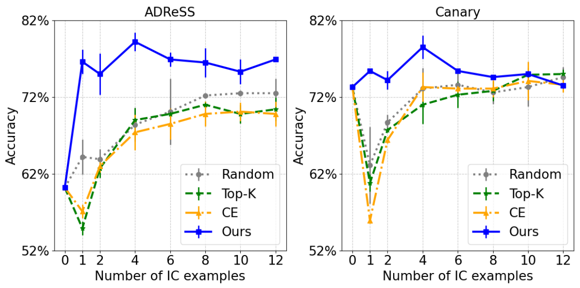

Figure 3: Impact of the number of in-context examples on ADReSS (left) and Canary (right) train sets.

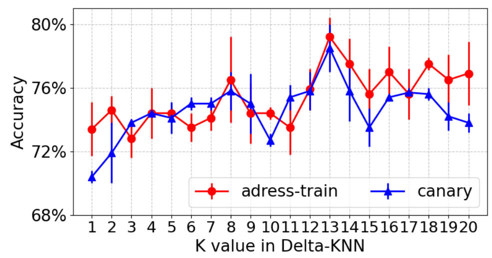

Figure 4: Impact of k value in Delta-KNN on ADReSS and Canary train sets with prompt (`R+C+L`; `G.CoT`).

2022). In contrast, Delta-KNN outperforms zero-shot, indicating its ability to select the most beneficial example (i.e., with the highest delta score) for the target input. When demonstrations include a balanced mix of positive and negative samples, Top-k, Random Sampling, and CE-based selection show improvements, particularly on ADReSS. However, on Canary, few-shot only begins win over zero-shot when $N \geqslant 4$. Across datasets, in-context performance increases, peaking at $N = 4$, after which it fluctuates and stabilizes. Thus, we select four in-context examples for our experiments.

5.4 Impact of k value in Delta-KNN

To evaluate the impact of k in Delta-KNN, we systematically vary k from 1 to 20 and perform cross-validation on the train sets of ADReSS and Canary. As shown in Figure 4, performance initially improves as k increases, then declines and fluctuates. Empirically, we find that $k = 13$ yields the best results on both datasets, so we adopt this value.

We further examine the effect of k under different prompts and observe varying optimal values. This suggests that determining the optimal number of target "representatives" in the Delta Matrix is non-trivial, as it possibly depends on multiple factors, including the prompt, language model, similarity computation, and text embedding model. As a result, determining the best k requires a case-by-case approach. For this reason, we rely on a held-out training set to empirically identify the best k. In future work, we aim to develop more advanced methods for optimizing this hyperparameter.

5.5 Delta-KNN using Other Text Encoders

Beyond OpenAI embeddings, we investigate LLM hidden states as text representations, based on the assumption that the same LLM can better capture subtle linguistic nuances. We perform experiment with Llama-3.1-8B-Instruct using two common strategies: extracting embedding of an appended [EOS] token at the end of the text (Wang et al., 2023a) and computing mean-pooled hidden states. Both approaches are applied at the first (L0), middle (L8, L16, L24), and the final layer (L32).

Figure 5 presents the 4-shot ICL results on ADReSS-train using three encoding methods: OpenAI embeddings, [EOS] token and mean-pooled hidden states. Surprisingly, LLM-derived embeddings do not outperform external embeddings, with the best [EOS] and mean-pooled representations achieving 74.5% and 77.3% accuracy, respectively. Comparing the two approaches, we observe that mean-pooling provides more stable performance, while [EOS] embedding shows greater variance across different layers. The choice of layer also significantly impacts performance: mid-layers such as L16 and L24 outperform the last layer (L32), suggesting that mid layers encode richer semantic meaning, which is in line with Chuang et al. (2023). Presumably, a single layer's hidden states may capture only limited aspects of the input text. Future work could explore combining representations from multiple layers to enhance text encoding (Li et al., 2025). Additionally, we note recent advancements in transforming LLMs into effective text encoders, such as LLM2Vec (BehnamGhader et al., 2024). Applying these methods could further boost the performance of Delta-KNN.

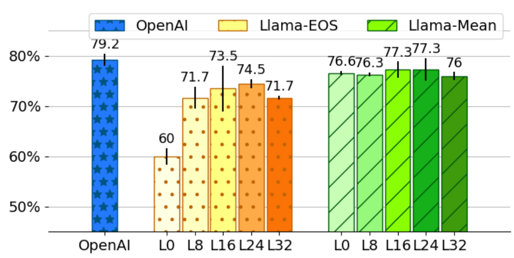

Figure 5: ADReSS-train performance using OpenAI `embedding-3-large` *vs.* Llama-3.1-8B hidden states ([EOS] and mean-pooling) over different layers.

	ADReSS-train	ADReSS-test	Canary
Mistral-7B-Instruct-v0.3			
Zero-shot	$52.3_{0.5}$	$67.7_{1.0}$	$63.1_{0.8}$
Random	$62.0_{2.8}$	$70.8_{2.1}$	$55.0_{0.4}$
Top-k Select.	$53.2_{2.3}$	$63.5_{3.1}$	$62.3_{0.0}$
CE Select.	$61.1_{1.9}$	$66.7_{4.2}$	$58.8_{3.5}$
Ours	$\mathbf{69.9_{1.4}}$	$\mathbf{76.0_{5.2}}$	$\mathbf{72.3_{0.4}}$
Qwen2.5-7B-Instruct			
Zero-shot	$61.6_{0.5}$	$66.8_{2.2}$	$63.5_{0.4}$
Random	$62.0_{2.8}$	$57.3_{1.0}$	$64.6_{3.8}$
Top-k Select.	$58.8_{1.4}$	$66.7_{2.1}$	$53.1_{6.2}$
CE Select.	$58.8_{0.5}$	$65.8_{5.3}$	$60.0_{1.5}$
Ours	$\mathbf{63.4_{0.5}}$	$\mathbf{67.7_{0.0}}$	$\mathbf{66.1_{2.7}}$

Table 3: AD detection accuracy using Mistral and Qwen LLMs, with prompt (`Role+Context+Linguistic`; `G.-CoT`). The best score within each LLM is in **bold**.

	ADReSS-train	ADReSS-test	Canary
Statistical ML Classifiers			
SVM (2021)	80.7	79.9	$51.9_{3.5}$
NN (2021)	76.2	77.1	-
RF (2021)	73.8	75.7	$68.7_{1.9}$
LR (2021)	-	-	$69.2_{1.4}$
Transfer learning-based PLM			
BERT (2021)	$\mathbf{81.2_{1.9}}$ (*81.8)	$79.3_{3.2}$ (*83.3)	$71.7_{2.6}$
GPT-3+SVM (2022)	80.9	80.3	-
Fine-tuned LLM			
Llama-3.1-8B	$70.8_{2.3}$	$77.1_{0.1}$	$63.8_{4.1}$
Delta-KNN ICL			
Ours (Llama)	$80.0_{1.3}$	$\mathbf{83.6_{2.0}}$	$\mathbf{78.5_{1.5}}$

Table 4: Accuracy using supervised baselines *vs.* ours. On Canary, we re-implement SVM, RF, and LR following Jang et al. (2021). We fine-tune BERT (with (*) scores directly from Balagopalan et al. (2021)) and Llama. Best score per column is in **bold**.

5.6 Delta-KNN with Other LLMs

We test the robustness of Delta-KNN on Mistral-7B-Instruct-v0.3 and Qwen2.5-7B-Instruct models. The results in Table 3 demonstrate that our method consistently outperforms other demonstration selection baselines across all tested LLMs, with Llama achieving the highest overall performance on both datasets. A closer analysis of performance variations across different prompts reveals that LLMs respond differently to the same instructions (detailed scores in Appendix F). In essence, Llama and Mistral perform best when provided with comprehensive prompts that include complete background information (`Role+Context+Linguistic`) and encourage step-by-step reasoning before making a prediction (`CoT`). In contrast, Qwen achieves its highest accuracy when prompted for a direct answer without explicit reasoning (`Role+Linguistic`). Interestingly, other demonstration selection methods also experience performance drops on Qwen when used more complex prompts, suggesting that prompt effectiveness is model-dependent. However, our approach remains robust and consistently improves performance across different prompting scenarios.

5.7 Delta-KNN *vs.* Supervised Baselines

Finally, we benchmark our Delta-KNN ICL with supervised baselines, see results in Table 4. Traditional supervised methods, such as statistical machine learning and transfer-learning approaches, achieve strong results. However, fine-tuning LLMs on this task does not lead to performance gains and instead underperforms compared to smaller supervised classifiers. This is expected, as our extremely small dataset likely lacks diversity and leads to overfitting (Garcia et al., 2023). This finding aligns with Vieira et al. (2024), which shows that fine-tuning Llama on limited datasets ($1k$ samples) can degrade performance. For ICL, we evaluate Delta-KNN using model-optimized prompts, i.e., the best-performing prompt for each dataset, as shown in §5.2. Excitingly, our approach achieves a new SOTA accuracy of 78.5% on Canary while delivering competitive performance on ADReSS.

Beyond strong performance, LLMs offers additional value by providing interpretable explanations that can assist doctors in diagnosis. To explore this, we conduct a qualitative study, where clinicians in our group compare LLM predictions with their own notes from a subset in Canary (Appendix G). Our findings suggest that LLMs strictly follow instructions and provide structured and insightful analyses, complementing human diagnosis.

6 Conclusion

We investigate the potential of LLMs as health assistants for AD detection, focusing on enhancing ICL. To tackle with limited data and the complexity of the task, we propose a novel demonstration selection method based on empirical evidence to quantify relative gains and identify optimal examples. Extensive experiments show that our approach consistently outperforms existing baselines, achieving substantial gains, particularly on the more challenging Canary dataset. Moving forward, we intend to investigate alternative text encoding techniques and strategies for hyperparameter optimization. Intriguingly, our method can be easily adapted for other data-poor scenarios and future applications such as integration with multimodal foundation models.

Limitations

Constructing the Delta Matrix involves pairwise computations with a time complexity of $\mathcal{O}(n^2)$, where n represents the number of training examples. In practice, we leverage vLLM (`https://github.com/vllm-project/vllm`) for accelerated LLM inference. Given our small-data scenario, these computations remain feasible within standard computational resources. For prompts requiring only short answers (i.e., no CoT reasoning), inference for $10,000$ examples completes in approximately 10 minutes. Prompts incorporating CoT reasoning take around 1.5 hours. Notably, fine-tuning an LLM for just one epoch requires a similar runtime yet yields inferior results compared to our ICL approach. To scale our method to larger datasets, a possible solution is to apply clustering to the training examples, selecting a representative subset before constructing the Delta Matrix.

For nearest neighbor selection (KNN), we explore multiple approaches, utilizing both an external text encoder from OpenAI and LLMs' internal hidden states. Our findings indicate that using LLM's inner embeddings does not enhance performance. However, improvements in similarity computation could be achieved through learning a similarity metric via contrastive learning or adopting advanced techniques to transform LLMs into more effective text encoders, such as LLM2Vec. Further advancements in this point could also help in optimizing the hyperparameter k.

Finally, we evaluate Delta-KNN across three LLMs from different families to assess the robustness of our approach. We focus on small-to-mid size models (7B–8B), balancing computational efficiency with strong performance. Testing our method on larger models or state-of-the-art open-source reasoning models, such as DeepSeek-R1, is an exciting direction for future exploration.

Ethical Considerations

The diagnosis of neurodegenerative disease is complex and relies on many indices. Automatic AI systems could provide clinicians with further clues, possibly alleviating the need for the patients to go through expensive and invasive screening tests, but this is a long-reach goal. In the healthcare domain, there is a risk that AI-generated predictions or analyses may be misinterpreted or directly relied upon as expert diagnoses. We emphasize the need for caution in their use. It is clear that the systems developed **can not** substitute for a human expert, as a diagnosis is a medical act. Moreover, linguistic clues and reasoning generated from LLMs, while crucial, have to be interpreted together with patient's clinical notes.

We carefully select the datasets used in this study to minimize potential biases and ensure that no private information—such as participants' health, clinical, or demographic data—is disclosed. This is a main reason for us exclusively testing with open-source language models. As authorized member of DementiaBank, we strictly follow its usage guidelines and ethical considerations. For the Canary dataset, the data collection process received approval from the Clinical Research Ethics Board, and details regarding the Institutional Review Board (IRB) approval will be provided.

The conception, implementation, analysis, and interpretation of results were conducted solely by the authors without any AI assistance. We used ChatGPT to help us check the grammar during writing.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Felix Agbavor and Hualou Liang. 2022. Predicting dementia from spontaneous speech using large language models. *PLOS digital health*, 1(12):e0000168.

Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager, and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain*, 136(12):3727–3737.

Sharon Ash and Murray Grossman. 2015. Why study connected speech production. *Cognitive neuroscience of natural language use*, pages 29–58.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Aparna Balagopalan, Benjamin Eyre, Jessica Robin, Frank Rudzicz, and Jekaterina Novikova. 2021. Comparing pre-trained and feature-based models for prediction of alzheimer's disease based on speech. *Frontiers in aging neuroscience*, 13:635945.

Oswald Barral, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. 2020. Non-invasive classification of alzheimer's disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. In *Proceedings of the 1st Annual Conference on Language Modeling (COLM)*, Philadelphia, USA.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Cory Shravya Shetty, Y McLean, Greg Corrado, and Nicholas A Furlotte1 AndrewB Carroll. 2023. Multimodal llms for health grounded in individual-specific data. In *Machine Learning for Multimodal Healthcare Data: First International Workshop, ML4MHD 2023, Honolulu, Hawaii, USA, July 29, 2023, Proceedings*, volume 14315, page 86. Springer Nature.

Mondher Bouazizi, Chuheng Zheng, Siyuan Yang, and Tomoaki Ohtsuki. 2023. Dementia detection from speech: what if language models are not the answer? *Information*, 15(1):2.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Hejie Cui, Zhuocheng Shen, Jieyu Zhang, Hui Shao, Lianhui Qin, Joyce C Ho, and Carl Yang. 2024. Llms-based few-shot disease predictions using ehr: A novel approach combining predictive agent reasoning and critical agent instruction. *arXiv preprint arXiv:2403.15464*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2024. A bibliometric review of large language models research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology*, 15(5):1–25.

Anna Favaro, Najim Dehak, Thomas Thebaud, Jesús Villalba, Esther Oh, and Laureano Moro-Velázquez. 2024. Discovering invariant patterns of cognitive decline via an automated analysis of the cookie thief picture description task. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pages 201–208.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences*, 26:243–254.

Kathleen C Fraser, Kristina Lundholm Fors, Marie Eckerström, Fredrik Öhman, and Dimitrios Kokkinakis. 2019. Predicting mci status from multimodal language data using cascaded classifiers. *Frontiers in aging neuroscience*, 11:205.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Julian Fritsch, Sebastian Wankerl, and Elmar Nöth. 2019. Automatic diagnosis of alzheimer's disease using neural network language models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5841–5845. IEEE.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10867–10878.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162, Singapore. Association for Computational Linguistics.

Hyeju Jang, Thomas Soroski, Matteo Rizzo, Oswald Barral, Anuj Harisinghani, Sally Newton-Mason, Saffrin Granby, Thiago Monnerat Stutz da Cunha Vasco, Caitlin Lewis, Pavan Tutt, et al. 2021. Classification of alzheimer's disease leveraging multi-task machine learning analysis of speech and eye-movement data. *Frontiers in Human Neuroscience*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, et al. 2024. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*.

Ehsan Kamalloo, Xinyu Zhang, Odunayo Ogundepo, Nandan Thakur, David Alfonso-Hermelo, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Evaluating embedding apis for information retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 518–526.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Weirui Kong, Hyeju Jang, Giuseppe Carenini, and Thalia Field. 2019. A neural model for predicting dementia from language. In *Machine Learning for Healthcare Conference*, pages 270–286. PMLR.

Raymond Li, Yuxi Feng, Zhenan Fan, Giuseppe Carenini, Weiwei Zhang, Mohammadreza Pourreza, and Yong Zhang. 2025. DeTriever: Decoder-representation-based retriever for improving NL2SQL in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8173–8183, Abu Dhabi, UAE. Association for Computational Linguistics.

Sheng-Chieh Lin, Minghan Li, and Jimmy Lin. 2023. Aggretriever: A simple approach to aggregate textual representations for robust dense passage retrieval. *Transactions of the Association for Computational Linguistics*, 11:436–452.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. In *INTERSPEECH 2021*. ISCA.

Xueguang Ma, Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2022. Document expansion baselines and learned sparse lexical representations for ms marco v1 and v2. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3187–3197.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. In *BioNLP 2017*, pages 232–237.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023b. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.

Shuyin Ouyang, Jie M Zhang, Mark Harman, and Meng Wang. 2023. Llm is like a box of chocolates: the non-determinism of chatgpt in code generation. *arXiv preprint arXiv:2308.02828*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Roy H Perlis. 2023. Application of gpt-4 to select next-step antidepressant treatment in major depression. *MedRxiv*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Claudia Rodríguez-Aranda, Stein Harald Johnsen, Petter Eldevik, Sigurd Sparr, Gry C Wikran, Marit Herder, and Torgil Riise Vangberg. 2016. Neuroanatomical correlates of verbal fluency in early alzheimer's disease and normal aging. *Brain and language*, 155:24–35.

Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J Nestor. 2012. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, 26(6):847–866.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2024. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations (ICLR)*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023a. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Zhuo Wang, Rongzhen Li, Bowen Dong, Jie Wang, Xiuxing Li, Ning Liu, Chenhui Mao, Wei Zhang, Liling Dong, Jing Gao, et al. 2023b. Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today. *arXiv preprint arXiv:2306.01499*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Jasper Xian, Tommaso Teofili, Ronak Pradeep, and Jimmy Lin. 2024. Vector search with openai embeddings: Lucene is all you need. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1090–1093.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Kathryn M Yorkston and David R Beukelman. 1980. An analysis of connected speech samples of aphasic and normal speakers. *Journal of speech and hearing disorders*, 45(1):27–36.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Youxiang Zhu, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021. Exploring deep transfer learning techniques for alzheimer's dementia detection. *Frontiers in computer science*, 3:624683.

A Data Statistics and Preprocessing

Statistics. Table 5 shows the length (average and standard deviation of number of tokens, tokenized by BERT model), demographic (age, gender) and clinical (cognitive tests) information on ADReSS and Canary datasets.

Analyzing document length, we observe that healthy controls generally produce longer speech compared to AD patients, with more detailed descriptions and longer sentences. However, documents in the Canary dataset are significantly longer than those in ADReSS and exhibit greater variation in length. The large variability suggests that Canary presents a more challenging dataset for AD detection.

Pre-processing. For ADReSS dataset, we extract clean texts by removing interviewer's content and special tokens such as non-verbal sounds encoded in the CHAT (Codes for Human Analysis of Transcripts) format. We only use the textual transcripts.

For Canary dataset, participants completed four tasks–pupil calibration, picture description, paragraph reading, and memory recall–during which both language and eye movement data were collected. In this study, we only utilize data from the picture description task. we use WhisperX (Bain et al., 2023) to first automatically transcribe the original speech data. The transcripts are manually verified by a human annotator to correct word spellings and speaker diarization.

B Supervised Classifiers

Following Jang et al. (2021), we re-implement the supervised methods using Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF), all implemented with the Scikit-learn library (Pedregosa et al., 2011). To ensure robust evaluation, we perform 10-fold cross-validation using ten different random seeds. The average scores are given in Table 6, in comparison with our Delta-KNN results.

	ADReSS		Canary	
	AD	Control	AD	Control
Training				
# Doc	54	54	63	67
Avg. # Token	122.2	134.9	150.9	206.2
Std. # Token	76.2	85.2	102.5	156.4
Age	66.7 ± 6.6	66.4 ± 6.5	72 ± 9	62 ± 15
Gender	24M / 30F	24M / 30F	31M / 34F	22M / 45F
Cognitive	17.1 ± 5.5	29.1 ± 1.0	18 ± 7	27 ± 3
Test				
# Doc	24	24	-	-
Avg. # Token	115.8	154.9	-	-
Std. # Token	66.2	107.6	-	-
Age	66.1 ± 7.4	66.1 ± 7.1	-	-
Gender	13M / 11F	13M / 11F	-	-
Cognitive	19.5 ± 5.4	28.9 ± 1.5	-	-

Table 5: Dataset demographic and clinical statistics. On cognitive tests, ADReSS reports Mini-Mental Status Examination score (MMSE); Canary reports Montreal Cognitive Assessment score (MoCA). - not applicable.

Model	ACC	AUC	SEN	SPE
SVM	51.9 ± 3.5	43.3 ± 6.8	25.2 ± 10.8	79.7 ± 5.4
RF	68.7 ± 1.9	73.6 ± 1.9	67.0 ± 2.9	70.2 ± 4.0
LR	69.2 ± 1.4	73.6 ± 1.4	69.9 ± 1.0	68.3 ± 2.3
Ours	78.5 ± 1.5	79.8 ± 0.9	70.6 ± 0.8	85.8 ± 2.2

Table 6: Comparison of supervised classifiers (top) and our Delta-KNN ICL approach with Llama (bottom). RF: random forest, LR: logistic regression. Supervised results are averaged over 10-seed 10-fold cross-validation.

Note that noting that our results differ slightly from those reported in Jang et al. (2021), as we do not use the exact same training samples (79 Patients and 83 Controls *vs.* our dataset with 63 Patients and 67 Controls). Additionally, we employ different speech-to-text methods, which may have led to variations in the transcripts.

C BERT Fine-tuning

We fine-tune BERT on ADReSS-train and Canary following Balagopalan et al. (2021), using the `bert-base-uncased` model (Devlin, 2018). We use the [CLS] token from the final hidden state as the aggregate representation and pass it to the classification layer. BERT model is fine-tuned for 10 epochs with a learning rate of $2e-5$, the same as in Balagopalan et al. (2021). Scores in Table 4 regarding BERT are averaged over five runs.

D Llama Fine-tuning

We also explore the feasibility of fine-tuning LLMs directly on our datasets. Given the extremely small size of our training data, we adopt LoRA (Hu et al.,

Hyperparameters	Selected
BitsAndBytes Quantisation	
use_4bit_quantization	True
use_nested_quant	True
bnb_4bit_compute_dtype	bfloat16
PEFT LoRA	
Lora r	8
Lora alpha	16
Lora dropout rate	0.1
Bias	none
Task type	CAUSAL_LM
Target modules	q_proj,k_proj,v_proj,o_proj
Training Arguments	
Training epoch	1
Batch size	1
Optimizer	adam
Learning rate	$1e-4$
Learning rate scheduler	cosine
Warm-up ratio	0.0
Weight decay	$1e-4$

Table 7: Hyperparameters for Llama-3.1-8B-Instruct fine-tuning.

2022), a parameter-efficient fine-tuning approach. Specifically, we use low-rank ($r = 8$) and low-alpha ($\alpha = 16$) values while restricting updates to attention modules (Q, K, V, O) to mitigate overfitting.

We fine-tune for a single epoch, as the training loss converges well, while additional epochs lead to a rebound in validation loss, indicating overfitting. All experiments are conducted on a single NVIDIA A100 40G GPU.

For hyperparameter selection, we tested multiple configurations, including different rank values (8, 16), alpha values (16, 32), and target modules (“*all-linear*”, “*q_proj,v_proj*”, “*q_proj,k_proj,v_proj,o_proj*”). Our results show that using lower alpha and dropout rates, combined with attention-only target modules, yields the best performance. Detailed values for hyperparameters are presented in Table 7.

E Prompt Templates

We provide prompting template used in our experiments in Table 9 on the next page.

F Results with Mistral and Qwen

We present AD prediction results using Mistral and Qwen with the prompt (`Role+Context+Linguistic; Demonstrations; G.-CoT`) in Table 10, while results with other prompts are shown in Table 11, both on page 16.

As discussed in §5.6, different LLMs respond differently to the same prompt. Generally, more comprehensive prompts tend to yield better performance, as observed with Llama and Mistral. However, Qwen performs better with a simpler prompt. As shown in the last section of Table 11, Qwen achieves its highest accuracy when provided with `Role+Linguistic` and no CoT reasoning (prompt 3). When additional background information and CoT reasoning are introduced, its performance declines across all demonstration selection methods.

	Diagnosis	MoCA	Prediction	
			LLM	Human
Case 1	AD	3	P	P
Case 2	Mild-moderate AD	16	P	P
Case 3	AD	16	P	P
Case 4	Mild AD	25	P	H
Case 5	aMCI*	27	P	P
Case 6	Healthy control	21	H	P
Case 7	Healthy control	25	P	H
Case 8	Healthy control	28	H	H
Case 9	Healthy control	29	H	H
Case 10	Healthy control	30	H	H

Table 8: Llama and human prediction on ten cases in Canary. aMCI*: Amnestic mild cognitive impairment (aMCI). Prediction highlighted in red is incorrect.

G Case Study on LLM’s Prediction

We conduct a qualitative study to examine how LLM-generated diagnoses compare with those made by a clinician in our research group. Specifically, we ask the clinician to provide diagnoses and reasoning for ten participants based solely on their picture description task outputs—without access to clinical notes—using similar instructions given to the LLM, see “Instruction to human” in Table 12 and Table 13.

Table 8 shows the predictions made by both the clinician and the LLM, alongside the ground-truth diagnoses and each participant’s Montreal Cognitive Assessment (MoCA) score. In this evaluation, the clinician correctly diagnosed eight cases, while Llama, utilizing Delta-KNN ICL, correctly identified nine. To illustrate the comparison in greater detail, we present the predictions and analyses from both the LLM and the clinician for Case 4 (Table 12) and Case 7 (Table 13).

In both cases, the clinician diagnosed the sub-

	Template
Background Prompt	`Role:` You are a medical expert in Alzheimer's disease. `Context:` The Boston Cookie Theft picture description task is a well established speech assessment in Alzheimer's disease. During the task, participants are shown the picture and are asked to describe everything they see in the scene using as much time as they would like. The objects (also known as information units) in this picture includes: "cookie", "girl", "boy", "woman", "jar", "stool", "plate", "dishcloth", "water", "window", "cupboard", "curtain", "dishes", "sink". `Linguistic:` You analyze linguistic features in the patient's speech, such as lexical richness, syntactic complexity, grammatical correctness, information units, and semantic coherence. Based on the participant's description of the picture, provide an initial diagnosis of dementia patient (P) and healthy control (H).
Example Prompt	`Zero-shot: None` `Demonstration:` Example: ## Text: <text> ## Answer: healthy control (H). ## Text: <text> ## Answer: dementia patient (P).
Question Prompt	`CoT:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). First explain step-by-step and then give a prediction with a probability. `Guided CoT:` Given the text below, classify the participant as a dementia patient (P) or healthy control (H). Please first reason from the following perspectives: (1) Vocabulary richness: such as the usage of different words; (2) Syntactic complexity: such as the length of the sentence and the number of subordinate clauses; (3) Information content: whether the participant describe most of the information units in the picture; (4) Semantic coherence: such as the usage of connectives and the change in description from one information unit to another; (5) Fluency and repetitiveness: whether the text is fluent with less repetitive sentences. Based on your reasoning, please give a prediction and the corresponding probability.

Table 9: Prompt template used for AD detection.

jects as healthy controls, whereas Llama predicted them as patients. A closer analysis reveals that Llama follows a strictly structured approach by sequentially analyzing the input according to the Guided Chain-of-Thought prompt (G.-CoT, shown in Table 9) before summarizing its findings. In contrast, the clinician relies on pragmatic considerations, focusing on higher-level cognitive markers such as *inference*, *causality statements*, and *logical event sequences*. However, this approach appears to overlook lower-level linguistic cues, such as lexical and syntactic patterns.

The clinician's diagnostic approach aligns more closely with human reasoning, as it highlights aspects that may be particularly revealing in assessing AD. Meanwhile, Llama's analysis is systematic and precise, offering high readability and interpretability. Its diagnosis is directly rooted in the input text, providing detailed explanations for each aspect. For instance, it explicitly points out structural errors, such as: "There are some errors in sentence structure, such as 'And his mother is not really thinking about washing up because the water is running over the sink.'." This level of detailed reasoning and explanation could be valuable in assisting clinicians by offering an additional layer of linguistic analysis that might otherwise be overlooked.

Method	ADReSS-train ACC	AUC	SEN	SPE	ADReSS-test ACC	AUC	SEN	SPE	Canary ACC	AUC	SEN	SPE
Llama-3.1-8B-Instruct												
Zero-shot	$62.2_{0.0}$	$60.1_{0.0}$	$98.1_{0.0}$	$22.2_{0.0}$	$57.6_{1.0}$	$57.6_{1.0}$	$100.0_{0.0}$	$15.3_{2.0}$	$73.3_{0.4}$	$72.1_{1.0}$	$79.4_{0.0}$	$67.7_{0.7}$
Random	$68.4_{2.2}$	$71.9_{3.1}$	$84.0_{2.3}$	$48.8_{6.3}$	$75.7_{4.3}$	$81.5_{2.6}$	$93.1_{2.0}$	$58.3_{9.0}$	$73.1_{2.7}$	$75.3_{3.7}$	$72.0_{3.3}$	$74.1_{2.5}$
Top-k Select.	$69.0_{1.6}$	$71.9_{2.5}$	$88.3_{2.3}$	$45.7_{1.7}$	$70.1_{2.0}$	$80.0_{0.8}$	$91.7_{3.4}$	$48.6_{2.0}$	$71.0_{2.5}$	$75.0_{2.2}$	$76.7_{0.7}$	$65.7_{4.2}$
CE* Select.	$67.4_{2.3}$	$74.5_{1.3}$	$85.2_{1.5}$	$45.7_{3.1}$	$70.1_{1.0}$	$76.4_{2.6}$	$93.1_{2.0}$	$47.2_{2.0}$	$73.3_{1.9}$	$78.4_{0.9}$	$79.9_{2.0}$	$67.2_{4.4}$
Delta-KNN (ours)	<u>**$79.2_{1.2}$**</u>	$78.9_{1.3}$	$69.1_{0.9}$	$85.2_{1.5}$	<u>**$80.5_{3.9}$**</u>	$85.8_{0.9}$	$70.8_{5.9}$	$86.1_{2.0}$	<u>**$78.5_{1.5}$**</u>	$79.8_{0.9}$	$70.6_{0.8}$	$85.8_{2.2}$
Mistral-7B-Instruct-v0.3												
Zero-shot	$52.3_{0.5}$	$61.5_{1.0}$	$94.4_{0.0}$	$10.2_{0.9}$	$67.7_{1.0}$	$76.3_{1.3}$	$100.0_{0.0}$	$35.4_{2.1}$	$63.1_{0.8}$	$65.3_{0.3}$	$77.8_{1.6}$	$49.3_{0.0}$
Random	$60.2_{2.8}$	$69.1_{2.0}$	$91.7_{0.9}$	$28.7_{4.6}$	$70.8_{2.1}$	$78.5_{8.2}$	$91.7_{4.2}$	$50.0_{0.0}$	$55.0_{0.4}$	$58.6_{1.0}$	$79.4_{1.6}$	$32.1_{2.2}$
Top-k	$53.2_{2.3}$	$69.6_{4.2}$	$88.9_{5.6}$	$17.6_{0.9}$	$63.5_{3.1}$	$74.0_{1.6}$	$91.7_{0.0}$	$35.4_{6.2}$	$62.3_{0.0}$	$69.2_{1.9}$	$84.9_{2.4}$	$41.0_{2.2}$
CE* Select.	$61.1_{1.9}$	$78.0_{3.9}$	$93.5_{2.8}$	$28.7_{0.9}$	$66.7_{4.2}$	$78.0_{0.7}$	$97.9_{2.1}$	$35.4_{6.2}$	$58.8_{3.5}$	$64.2_{2.3}$	$84.9_{2.4}$	$34.3_{4.5}$
Delta-KNN (ours)	<u>$69.9_{1.4}$</u>	$82.4_{3.2}$	$90.7_{0.0}$	$49.1_{2.8}$	<u>$76.0_{5.2}$</u>	$84.9_{2.7}$	$95.8_{0.0}$	$56.2_{10.4}$	<u>$72.3_{0.4}$</u>	$74.8_{0.2}$	$86.5_{0.8}$	$59.0_{0.7}$
Qwen2.5-7B-Instruct												
Zero-shot	$61.6_{0.5}$	$64.9_{2.9}$	$94.4_{0.0}$	$28.7_{0.9}$	$66.8_{2.2}$	$65.5_{2.6}$	$97.9_{2.1}$	$43.8_{10.4}$	$63.5_{0.4}$	$62.5_{0.5}$	$69.0_{0.8}$	$58.2_{1.5}$
Random	$62.0_{2.8}$	$62.5_{2.2}$	$89.8_{2.8}$	$34.3_{2.8}$	$57.3_{1.0}$	$53.8_{0.7}$	$75.0_{4.2}$	$39.6_{2.1}$	$64.6_{3.8}$	$63.2_{3.8}$	$81.0_{4.8}$	$49.3_{3.0}$
Top-k Select.	$58.8_{1.4}$	$56.2_{0.0}$	$88.0_{2.8}$	$29.6_{0.0}$	$66.7_{2.1}$	$65.5_{6.0}$	$91.7_{0.0}$	$41.7_{4.2}$	$53.1_{6.2}$	$51.6_{7.1}$	$70.6_{5.6}$	$36.6_{6.7}$
CE* Select.	$58.8_{0.5}$	$58.9_{1.4}$	$88.0_{2.8}$	$29.6_{1.9}$	$65.8_{5.3}$	$63.8_{12.7}$	$91.7_{8.3}$	$45.8_{8.3}$	$60.0_{1.5}$	$57.4_{0.0}$	$68.3_{7.9}$	$52.2_{4.5}$
Delta-KNN (ours)	<u>$63.4_{0.5}$</u>	$62.7_{2.2}$	$82.4_{0.9}$	$44.4_{0.0}$	<u>$67.7_{0.0}$</u>	$62.2_{1.1}$	$85.4_{6.2}$	$47.9_{6.2}$	<u>$66.1_{2.7}$</u>	$64.8_{3.9}$	$71.4_{0.0}$	$45.5_{5.2}$

Table 10: AD detection results using different demonstration selection methods on Llama, Mistral, and Qwen models; prompt (`Role+Context+Linguistic; G.-CoT`). The best accuracy within each LLM is <u>underlined</u> while the overall highest accuracy is in **bold**.

	Role	Con.	Ling.	CoT	G.-CoT	ADReSS-train Delta-KNN (Rdm, Top$k$, CE)				ADReSS-test Delta-KNN (Rdm, Top$k$, CE)				Canary Delta-KNN (Rdm, Top$k$, CE)			
Llama-3.1-8B-Instruct																	
(1)	✗	✗	✗	✗	✗	73.0	↓ 13.9	↓ 17.6	↓ 12.7	69.8	↑ 0.3	↓ 2.8	↑ 0.8	63.1	↓ 2.3	↓ 2.3	↓ 3.9
(2)	✓	✓	✗	✗	✗	72.7	↓ 2.1	↓ 2.4	↓ 3.3	69.1	∼ 0	∼ 0	↓ 2.9	70.0	↓ 3.8	↓ 3.8	↓ 4.1
(3)	✓	✗	✓	✗	✗	73.1	↓ 7.8	↓ 13.1	↓ 6.9	74.4	↓ 5.3	↓ 2.6	↓ 1.2	68.1	↓ 9.1	↓ 2.2	↓ 2.7
(4)	✓	✓	✗	✓	✗	73.6	↓ 5.9	↓ 4.9	↓ 6.3	74.6	↓ 2.1	↓ 2.8	↓ 2.8	71.5	↓ 9.7	↓ 4.3	↓ 10.2
(5)	✓	✗	✓	✓	✗	74.5	↓ 10.2	↓ 14.5	↓ 16.3	74.6	↓ 11.1	↓ 13.2	↓ 16.7	65.1	↓ 3.6	↓ 6.9	↓ 7.7
(6)	✓	✓	✓	✓	✗	80.0	↓ 9.9	↓ 11.1	↓ 8.6	83.6	↓ 13.8	↓ 10.4	↓ 12.5	70.8	↓ 7.5	↓ 7.7	↓ 9.6
(7)	✓	✓	✓	✗	✓	79.2	↓ 10.8	↓ 10.2	↓ 11.8	80.5	↓ 2.8	↓ 8.4	↓ 8.4	78.5	↓ 5.4	↓ 7.5	↓ 5.2
Mistral-7B-Instruct-v0.3																	
(1)	✗	✗	✗	✗	✗	50.0	∼ 0	∼ 0	∼ 0	51.0	↓ 1.0	↓ 1.0	∼ 0	51.2	↓ 2.7	↓ 1.5	↓ 1.2
(2)	✓	✓	✗	✗	✗	51.4	↓ 1.4	↓ 1.4	↓ 1.4	53.1	↓ 0.9	↓ 2.1	↓ 3.1	55.0	↓ 6.5	↓ 6.5	↓ 6.5
(3)	✓	✗	✓	✗	✗	50.5	↑ 0.4	↓ 0.5	↓ 0.5	55.2	↓ 4.2	↓ 5.2	↓ 5.2	49.6	↓ 0.4	↓ 0.9	↓ 0.9
(4)	✓	✓	✗	✓	✗	63.0	↓ 4.2	↓ 2.8	↓ 4.7	68.8	↓ 7.3	↓ 4.3	↓ 10.5	63.1	↓ 7.2	↓ 13.5	↓ 12.7
(5)	✓	✗	✓	✓	✗	58.8	∼ 0	↓ 9.7	↓ 1.4	65.6	↓ 4.1	↓ 10.4	∼ 0	65.4	↓ 10.2	↓ 9.8	↓ 6.2
(6)	✓	✓	✓	✓	✗	68.5	↓ 9.2	↓ 11.1	↓ 6.9	79.2	↓ 15.7	↓ 15.7	↓ 19.8	65.0	↓ 4.5	↓ 7.7	↓ 7.7
(7)	✓	✓	✓	✗	✓	69.9	↓ 9.7	↓ 16.7	↓ 8.7	76.0	↓ 5.2	↓ 12.5	↓ 9.3	72.3	↓ 17.3	↓ 10.0	↓ 13.5
Qwen2.5-7B-Instruct																	
(1)	✗	✗	✗	✗	✗	66.7	↓ 6.7	↓ 11.6	↓ 7.4	72.9	↓ 1.0	↓ 8.3	↓ 12.5	67.7	↓ 6.2	↓ 10.8	↓ 4.2
(2)	✓	✓	✗	✗	✗	70.8	↓ 7.2	↓ 11.5	↓ 6.0	71.9	↓ 6.3	∼ 0	↓ 6.3	67.3	↓ 4.2	↓ 3.1	↓ 0.5
(3)	✓	✗	✓	✗	✗	69.4	↓ 4.6	↓ 6.6	↓ 7.4	78.1	↓ 3.1	↓ 5.3	↓ 10.4	69.2	↓ 3.6	↓ 6.5	↓ 4.2
(4)	✓	✓	✗	✓	✗	70.0	↑ 1.8	↓ 4.7	↓ 1.9	69.7	↓ 7.2	↑ 1.2	↓ 4.2	61.9	↑ 0.4	↓ 1.5	↑ 0.6
(5)	✓	✗	✓	✓	✗	63.9	↓ 4.2	↓ 0.5	↓ 8.3	76.0	↓ 7.2	↓ 6.8	↓ 6.8	63.1	↑ 5.4	↓ 4.6	↓ 2.2
(6)	✓	✓	✓	✓	✗	59.7	↓ 2.6	↑ 2.8	↓ 0.9	64.6	↓ 3.1	↓ 2.1	↓ 0.9	56.9	↓ 8.7	↓ 4.0	↓ 4.5
(7)	✓	✓	✓	✗	✓	63.4	↓ 2.4	↓ 4.6	↓ 4.6	67.7	↓ 10.4	↓ 1.0	↓ 1.9	66.1	↓ 1.5	↓ 13.0	↓ 6.1

Table 11: Delta-KNN performance (accuracy) using different prompt engineering strategies (Role, Context, Linguistic cues, chain-of-thought reasoning (CoT), and Guided CoT) on ADReSS and Canary datasets, in comparison with Random sampling (Rdm), Top-k (Liu et al., 2022) selection, and Conditional Entropy (CE) (Peng et al., 2024) selection. ↓, ↑, and ∼ symbols refer to lower, higher, and same accuracies compared to the Delta-KNN, respectively.

	Case study 4 (Gold label: P)
Instruction to Human	This experiment aims to explore how clinicians assess language in the Cookie Theft task and identify the key criteria (reasoning processes) they use to make judgments. Below is a transcript produced by a participant (either a dementia patient or a healthy control). Based on the participant's description of the picture, please provide a brief analysis (2-10 sentences). Then, make a prediction by classifying the participant as either a healthy control (H) or a dementia patient (P).
Instruction to LLM	See prompt template in Table 9.
Input Text	"There's a boy that's stealing cookies from the cookie jar. And his mother is not really thinking about washing up because the water is running over the sink. And the boy is going to fall because of the stool. And the only thing his sister is interested in is for him to get more cookies, throw more cookies around. And this looks like a driveway outside. And there's some cups and two cups and a plate. And I don't know, anything of note? No?"
Human Diagnosis	Healthy control. There were **inferences** made from the picture, such as identifying individuals and assuming their relationships (e.g., recognizing a mother, a boy, and a girl). Moreover, the participant was making **causality statements**, such as predicting that the boy is going to fall because of the unstable stool.
Llama Diagnosis	Here's the step-by-step analysis: **1. Lexical richness**: The participant uses a variety of words, including "stealing", "interested", and "driveway", which suggests a good vocabulary. **2. Syntactic complexity**: The participant's sentences are relatively simple, with a mix of short and longer sentences. However, there are some errors in sentence structure, such as "And his mother is not really thinking about washing up because the water is running over the sink." This sentence is a bit convoluted and could be rephrased more clearly. **3. Grammatical correctness**: The participant makes a few grammatical errors, such as using "And" instead of a more formal conjunction, and using a comma splice in the sentence "And the only thing his sister is interested in is for him to get more cookies, throw more cookies around." **4. Information units**: The participant mentions most of the objects in the picture, including the boy, mother, sister, cookie jar, stool, sink, cups, and plate. However, they seem to be having trouble with the window and the scene outside, which they incorrectly identify as a driveway. **5. Semantic coherence**: The participant's description is generally coherent, but there are some strange inferences, such as the sister being interested in the boy throwing cookies around. Based on these observations, I would classify the participant as a dementia patient (P) with a probability of 0.7. The participant's description is generally clear and coherent, but there are some errors in sentence structure, grammar, and semantic coherence that suggest cognitive impairment. However, the participant's vocabulary and ability to identify most of the objects in the picture suggest that they are not severely impaired.

Table 12: The first example of a comparison between Llama's and human's prediction on participant 4's speech. Human's prediction is incorrect, while Llama's prediction is correct.

	Case study 7 (Gold label: H)
Instruction to Human	This experiment aims to explore how clinicians assess language in the Cookie Theft task and identify the key criteria (reasoning processes) they use to make judgments. Below is a transcript produced by a participant (either a dementia patient or a healthy control). Based on the participant's description of the picture, please provide a brief analysis (2-10 sentences). Then, make a prediction by classifying the participant as either a healthy control (H) or a dementia patient (P).
Instruction to LLM	See prompt template in Table 9.
Input Text	"There is a boy trying to reach a cookie jar, falling off a stool almost. His sister reaching up for a cookie. His mother is washing the dishes. And the sink is overflowing with water. And I'm finished."
Human Diagnosis	Healthy control. This person was making lots of role assumptions, with **logical sequence** of events (from dishes to water overflowing).
Llama Diagnosis	Here's the step-by-step analysis: **1. Lexical richness**: The participant uses a limited vocabulary, with only 7 words or phrases to describe the scene. This is a characteristic of dementia patients, who often experience word-finding difficulties. **2. Syntactic complexity**: The participant's sentence structure is simple and lacks complexity. They use short, basic sentences with no subordinate clauses or embedded phrases. This is consistent with the language patterns of dementia patients. **3. Grammatical correctness**: The participant's sentences are grammatically correct, but there is a lack of detail and elaboration. This is a characteristic of dementia patients, who may struggle to provide nuanced descriptions. **4. Information units**: The participant identifies 5 key elements in the scene: the boy, the sister, the mother, the cookie jar, and the overflowing sink. However, they fail to mention other important details, such as the stool, the sister's actions, and the mother's apron. **5. Semantic coherence**: The participant's description is disjointed and lacks coherence. They jump abruptly from describing the boy's actions to stating that they are finished. This is a characteristic of dementia patients, who may struggle to maintain a coherent narrative. Based on these observations, I predict that the participant is a dementia patient (P) with a high probability of 0.85. The participant's language patterns, including their limited vocabulary, simple sentence structure, and lack of detail, are consistent with the characteristics of dementia patients.

Table 13: The second example of a comparison between Llama's and human's prediction on participant 7's speech. Human's prediction is correct, while Llama's prediction is incorrect.